# 基于大数据的股票价格预测研究案例教学指导手册

**教学适用的课程：**《大数据架构与设计》

**教学适用的专业：**电子信息

**教学目的与用途：**熟悉Hadoop、Spark等大数据技术栈；掌握分布式大数据处理环境的搭建；掌握大数据处理与分析的基本流程。

**教学内容:**

（1）理论依据：本案例基于大数据技术对股票价格进行建模与预测。案例整合了常用的数据挖掘与机器学习方法，进行数据爬取、数据预处理与预测，实现了股票价格走势的智能化预测，为股民的选择带来更有价值的参考。

（2）涉及知识点：Hadoop&Spark技术；分布式爬虫技术；数据预处理；回归预测。

（3）分析路径：首先在Hadoop&Spark基础环境下采用Scrapy+Redis+MongoDB的方式对股票数据进行爬取，其中Scrapy负责爬取内容信息，Redis搭建了一个分布式的缓存中间件，MongoDB负责数据的存取。在获得股票数据之后，首先对股票数据进行了清洗，并将这部分数据以7:3的比例划分为训练集和测试集。股票数据可以抽象为一维线性数据，对该类型的数据进行预测，一般会选用RNN或者LSTM，然而由于RNN在长期记忆方面表现得不太好，LSTM能够更好的记忆长期数据，能够得到更为准确的预测结果，因此在预测模型选取方面，选用了LSTM模型。训练模型时，考虑到本门课程中的分布式架构，在训练模型时选用了Elephas进行分布式的超参数优化，模型训练完成后使用该模型预测2020年之后的日线开盘价格，并与实际价格进行比对，最后采用Echarts进行数据的可视化。

**启发思考题:**

在进行股票价格预测的时候，为了避免模型与数据集之间的过拟合现象，你能想到哪些方法呢？（参考答案：采用正则化项和dropout机制）

**建议课堂计划：**

1. 时间安排：3课时
2. 学生学习准备：了解常用的大数据技术栈以及常见的数据预处理、回归预测方法的理论知识。
3. 分组及讨论内容：分组讨论启发思考题。
4. 案例开场白：股票交易作为一个充满博弈的系统，自诞生以来，就吸引了众多的统计学家对其进行预测研究，因而各种指标数据也层出不穷。随着深度学习的出现与流行，近年来出现了一些基于大数据以及机器学习、深度学习的预测系统。本节课将学习如何使用当下比较流行的LSTM模型对股票的开盘数据进行预测。
5. 结束总结：这节课我们在股票模型的预测任务上，使用了分布式爬虫技术，熟悉了LSTM模型，还引入了Elephas这一分布式训练的框架，最终实现对股票日线开盘价格数据的回归预测。我们通过平时课程所学知识去解决了实际社会生活中的问题，课后感兴趣的同学可以搜集相关资料，进一步了解大数据在其他行业的应用，相信你们会收获更多。
6. 案例引导建议：在教师简要分析应用场景后，励学生自主考虑如何使用所学知识解决该场景问题。

**参考文献:**

[1]吴宇鹏.分布式网络爬虫技术的研究与实现[J].电脑编程技巧与维护,2020(11):9-10+19.DOI:10.16184/j.cnki.comprg.2020.11.003.

[2]张胜敏,王爱菊.基于Python的分布式多主题网络爬虫的研究与设计[J].开封大学学报,2021,35(01):93-96.

[3]武博. 基于LSTM模型的股票价格预测[D].大连理工大学,2021.

[4]史国举.数据可视化技术在大数据分析领域的应用及发展研究[J].无线互联科技,2021,18(18):96-97.

**其他教学支持材料：**

1. 本案例提供配套的PPT、视频、数据集与代码等，发布于Github，链接为：https://github.com/Wanghui-Huang/CQU\_bigdata。

2. 本案例涉及到数据预处理以及多种机器学习算法，建议使用python语言进行编写，推荐的工具包有pandas（数据读取与预处理库），scikit-learn（机器学习算法库），Echarts（可视化绘图库）。